04: Categorical Variables + Intro to Quantitative Variables

Stat 120 | Fall 2025

Prof Amanda Luby

The lab manual data `Pew` contains data from a January 2014 Pew Research Center survey about the Internet. We'll consider two variables from this survey:

- `income`: the person's yearly income, grouped into four categories
- `values`: an answer to the question asking if the Internet has been a "good thing"

0. Import the data into R by running the following code chunk. Click on the dataset in the environment pane to pull up the data viewer. Can you see the "spreadsheet view"?

1. The code below computes the `frequency table` for the `values` variable. Add a "pipe" (`%>%`) and another command to answer: what proportion of respondents said that the Internet *has* been a good thing?

```
bad good
140  556
```

2. Using esquisse, make a bar plot of the counts of people who said the Internet has vs. has not been a good thing. Insert your code in the chunk below.

3. Show the two-way table for the `values` and `income` variables. How many people from the survey had an income above $150,000 a year and thought the Internet was a bad thing?

```
              bad good
0-30000        57  133
150000          5   64
30000-75000    50  206
75000-150000   28  153
```

4. The table in part 3 shows the income categories in the wrong order (i.e., not in order of increasing income). You can fix this by changing the order of the categories of the `income` variable as below (see Appendix A.1 of the lab manual). Remake the table after reordering the categories.

5. Use `esquisse` to make a *stacked barplot* of `income` and `values`.

6. Your bar plot by default should show the *counts* of each combination. Copy and paste your bar plot code, and change `geom_bar()` to `geom_bar(position = "fill")`. Explain what this chunk of code did.

Stop Here and let Amanda know you've finished with the categorical EDA section

A third variable in the dataset is `age`. The next few questions walk through how to do basic EDA for a quantitative variable.

7. Create a histogram of the `age` variable using esquisse

8. In the empty code chunk below, run `mean(pew$age)` and `median(pew$age)`. Now try `mode(pew$age)`. What happened?

9. Using your answers to the previous two questions, describe the distribution of `age`.